Kinected Conference: Augmenting Video Imaging with Calibrated Depth and Audio

Anthony DeVincenzi
MIT Media Lab
75 Amherst St
Cambridge, MA 02139
tonyd@media.mit.edu

Hiroshi Ishii
MIT Media Lab
75 Amherst St
Cambridge, MA 02139
ishii@media.mit.edu

Lining Yao
MIT Media Lab
75 Amherst St
Cambridge, MA 02139
liningy@media.mit.edu

Ramesh Raskar
MIT Media Lab
75 Amherst St
Cambridge, MA 02139
raskar@media.mit.edu

CSCW 2011, March 19–23, 2011, Hangzhou, China.
ACM 978-1-4503-0556-3/11/03.

Abstract
The proliferation of broadband and high-speed Internet access has, in general, democratized the ability to commonly engage in videoconference. However, current video systems do not meet their full potential, as they are restricted to a simple display of unintelligent 2D pixels. In this paper we present a system for enhancing distance-based communication by augmenting the traditional video conferencing system with additional attributes beyond two-dimensional video. We explore how expanding a system's understanding of spatially calibrated depth and audio alongside a live video stream can generate semantically rich three-dimensional pixels containing information regarding their material properties and location. We discuss specific scenarios that explore features such as synthetic refocusing, gesture activated privacy, and spatiotemporal graphic augmentation.

Keywords
Videoconferences, Stereo Camera, Synthetic Focusing, Augmented Reality

ACM Classification Keywords
H5.m. Information interfaces and presentation

General Terms
Design, Human Factors

Introduction

Traditional networked video systems have primarily focused on image clarity through pixel density and sensor quality, as well as real-time image transmission to reduce the illusion of latency. Though foundational in purpose, advancements in these components provide little capability beyond simplistic two-dimensional perception; the representation of people, places, and things is done so only in values of RGB, ignoring the many material and sensorial properties naturally perceived by the human eye. By attempting to capture, and embed the display pixels with rich data, such as depth and audio, we believe that live video systems can benefit from a perceptually natural, and computationally augmented experience.

For the first time, we have shown that by introducing spatially calibrated depth and audio with a networked video system, we are able to achieve a wide variety enhanced videoconference scenarios. We explore these scenarios through three cornerstone applications: synthetic focusing for enhanced depth perception, conversational augmentation through spatially contextual graphic images, and creating 'invisible privacy spaces' by altering the perception of "live space".

Related Work

Decades of research have gone underway towards expanding our ability to communicate remotely. The proliferation of broadband and high-speed Internet access has, in general, democratized the ability to commonly engage in videoconference. Historically, one of the largest challenges in remote conferences resides in simulating the perception of "seamlessness" between the two remote locations [6]. This phenomenon is easily disturbed not only through complications in data transmission, but also in a qualitative lack of "connectedness" experienced through a flat display. Research in distributed systems for enhanced video conferencing has been explored in research such as Open Shared Workspace [1], Augmented Reality Videoconferencing [2], and Psuedo-3D Video Conferencing [5]. We aim to stand upon this research and push further towards techniques for enhancing the remote video experience.

Previous research has also gone underway towards enhancing the perception of realism in remote video by incorporating three-dimensional depth information with volumetric, or holographic display [3,4]. These instances focus on superimposing three-dimensional imagery within the physical space, whereas our approach aims to embed the two-dimensional display with rich depth, audio, and semantic object information. In other words, we aim to exploit additional pixel data through a method that augments the videoconference experience, not intended to simulate volumetric representation.

Videoconference System	Computational Information	Depth & Focus
2D Camera	2D RGB Pixel Matrix	Infinity Focus, No Depth
3D Camera & Spatial Audio	3D RGB pixel matrix with depth and audio location.	Dynamic focus, Centimeter Accurate Depth

Table 1. Comparison of the capabilities of videoconference system with 2D camera and 3D camera

Design and Implementation

The system fundamentals consist of two networked locations, each containing a video screen for viewing the opposite space, a standard RGB digital web camera

enhanced by a depth sensing "3D camera" such as the Microsoft Kinect, and calibrated microphones for audio queue and location. Computational processing is applied to each of the video streams, incorporating simplistic forms of face tracking through openCV for human detection as well as a number of custom algorithms for the perceptual manipulation of space.

C++ and the openFrameworks library are used for video processing and effect rendering. Networking is handled explicitly through our application, but is intended for exploration purposes only.

Table 1 demonstrates that stereoscopic imaging calibrated with audio location can achieve robust computational results, notably in object focus. The combined depth and audio location provides a rich "spatial map" which is used to analyze and alter the representation of all pixels on a Cartesian axis.

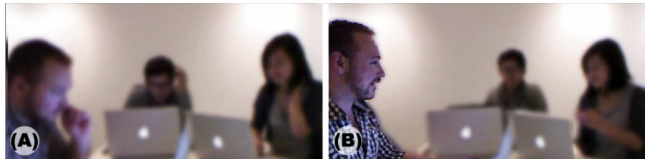

Figure 1: Synthetic focusing used to create a "Talk to Focus" effect. Background blur is generated based on depth map

Scenario 1: Synthetic focusing

In visual conversation we commonly adjust the focus of our gaze towards those who are speaking. Without conscious thought, our eyes dart in and out of different focal depths, expanding and contracting the aperture of our inner eye. The subtle depth-of-field created when focusing on who you are speaking with is a natural tool which affords literal and cognitive focus. In the case of videoconferencing, this natural trait is reduced by the inherent constraints of a "flat screen". Because our eyes cannot focus based on distance, vision is reduced to the cold and impersonal full-focus that nearly all webcams provide. We propose a scenario where subjects are taken in and out of focus computationally, with no need for additional lens or mechanical parts. Referred to as *Talk to Focus*, the system recognizes those currently speaking and places them within a blur-free range of the depth field. Due to our ability to infer many different layers of depth in a single scene, multiple areas of focus and soft transitions in focal range can be simulated (figure 1).

Figure 2: Gesture activated privacy demonstrated by a) the actual scenario captured by an ordinary camera b) live video stream as seen by remote viewer. Red box demonstrates privacy zone activated by right occupant with frozen pixels.

Scenario 2: Gesture Activated Privacy

In contrast to our first scenario, where we are simulating natural human perception, our second application attempts the opposite by allowing the user to render oneself, or specified area, invisible with a gestural command (figure 2). This technique is useful for executing small tasks such as checking email, short conversation, or temporarily leaving the room when it may be considered rude as viewed from the remote location. By freezing specific pixels at a certain depth, the simulation does not interrupt objects moving in the foreground of where time has been frozen. Similarly, privacy zones can be designated with a number of parameters such as the ability to be rendered invisible, have your face hidden, or the ability to remove all direct audio from the specific location.

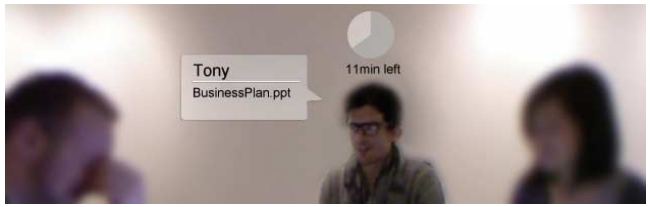

Figure 3: Contextual bubbles showing speaker's information such as name, shared documents, speaking time, ect.

Scenario 3: Spatial Graphic Augmentation

Our third and final scenario exemplifies the traditional use of augmented reality by associating dynamic context tags with attendees within the remote location (figure 3). Tags are virtually rendered in dimensional space alongside their host. This is achieved by combining video depth, audio location, and facial tracking algorithms in such a way that most accurately represents the place-in-space where an actual object, or tag would appear. Tags contain information relevant to their host such as name, location, shared files during a conference, and the ability to see time specific attributes such as total time or time since talking. Tags can be interpreted as a metaphor for representing a person as an interactive object, where the tag itself may not be literal, and could reveal itself in many different forms: either by audio queued automation or clicking a cursor on the remote person to reveal the additional graphic and textual information.

Conclusion

In lieu of overabundant research focused solely on volumetric and holographic display, we believe that the Depth in Conversation system reveals a largely unexplored space in augmenting the experience captured within current display technology. Having access to dynamic information regarding distant spaces, such as calibrated depth and audio, provides us with a compelling framework, further enhancing the experience of distributed video communication. We present our system as a set of tools for exploring both the subtle and more evident methods of augmented video imaging.

References

[1] Hiroshi Ishii and Naomi Miyake. 1991. Toward an open shared workspace: computer and video fusion approach of TeamWorkStation. *Commun. ACM* 34, 12 (December 1991), 37-50.

[2] Istvan Barakonyi, Tamer Fahmy, and Dieter Schmalstieg. 2004. Remote collaboration using Augmented Reality Videoconferencing. In *Proceedings of Graphics Interface 2004* (GI '04). Canadian Human-Computer Communications Society, School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, 89-96.

[3] Ramesh Raskar, Greg Welch, Matt Cutts, Adam Lake, Lev Stesin, and Henry Fuchs. 1998. The office of the future: a unified approach to image-based modeling and spatially immersive displays. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques* (SIGGRAPH '98). ACM, New York, NY, USA, 179-188.

[4] Henry Fuchs, Gary Bishop, Kevin Arthur, Leonard McMillan, Ruzena Bajcsy, Sang Lee, Hany Farid, and Takeo Kanade. 1994. *Virtual Space Teleconferencing Using a Sea of Cameras*. Technical Report. University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

[5] Chris Harrison and Scott E. Hudson. 2008. Pseudo-3D Video Conferencing with a Generic Webcam. In *Proceedings of the 2008 Tenth IEEE International Symposium on Multimedia* (ISM '08). IEEE Computer Society, Washington, DC, USA, 236-241.

[6] Martin Kuechler and Andreas M. Kunz. 2010. Collaboard: a remote collaboration groupware device featuring an embodiment-enriched shared workspace. In *Proceedings of the 16th ACM international conference on Supporting group work* (GROUP '10). ACM, New York, NY, USA, 211-21